2006 年度卒業研究概要

「タグ」の関連構造による集合知の整理手法

大谷　紀子研究室

0332177　二階堂　了太

1．研究背景と目的

タグとは一塊の情報に対して自由につけられる簡潔な語や句を指し、図 1 のように一つの情報に対して複数の人間が同じタグを付けたり、別々の情報に対して同じタグを付与したりする場合がある。長い文章、画像や音楽といったデータ量の大きい情報に、短いテキストデータであるタグをいくつか付与しておくことで、情報の中身を確認することなく、小さな労力で重要さや内容の範疇などが判断できる。

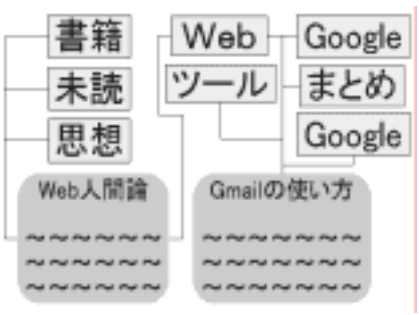

図 1：情報に対するタグの付き方

blog やソーシャルブックマーキングなど、近年増加したユーザ参加型の Web サービスで広くタグが用いられている。実質上、単なるテキストデータとして扱えるため可搬性が高く、異質なシステムを横断してサービス同士をつなぐために用いられる場合もある。

各種サービスのユーザ数の増加に伴いタグ自体も増加し続けている。タグはユーザによって様々な利用がなされるため、タグの単純な増加は有益なデータのみではなくノイズとなるデータの増加にもつながり、利用者一人一人の利便性や情報の濃度が損なわれていく面もある。

本研究は、タグと情報の関連構造を元にしたクラスタ分析によって類似のタグ同士を発見、整理し、タグデータを凝縮して、より利便性の高いものにすることを目的とする。

2．タグの整理手法

タグの関連構造とは情報とタグの結びつきを一つの構造として見たもので、情報に対するタグの付き方によって変わる。関連構造をイメージで捉えると図 2 のようになり、似た構造を取る部分の多さによってタグとタグの類似度を判断できると考えられる。

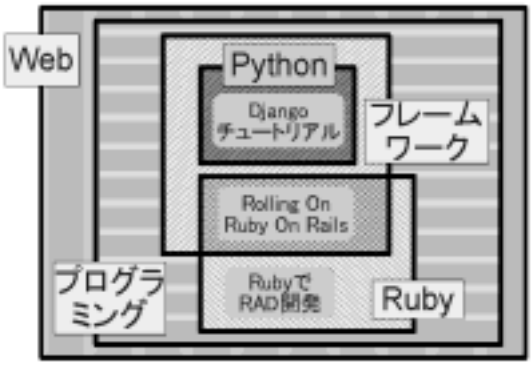

図 2：タグの関連構造

本研究では、はてなブックマーク[1]から、ユーザの被参照数が比較的多いブックマーク情報をランダムに 299 件を取得し、付加されている 3840 種 40667 個のタグを対象に、タグの関連構造を数種の凝集型の階層的クラスタ分析によって処理し、結果を再度組み合わせて類似のタグ同士をまとめ、整理した。

タグの中には、利用者個人にのみ帰属し大多数のユーザには全く無意味なものや、同じ意味を持つが異なる表現であるものなどが存在する。類似タグを事前に求めておくことで、不要なタグを一括で無視することや、ほぼ同一文脈に用いられるタグ同士を同一視することが可能となり、タグを利用した行為の効率化を図ることができる。

3. クラスタリング

階層的クラスタ分析では結果として、図 3 のようなタグを葉とした樹形の構造が得られ、タグ同士の距離を確認できる。中でも周囲 4〜6 個程度のタグの関連性が非常に高く、逆に周囲 4〜6 個よりも離れた距離を持つタグ同士は関連性があるとはいえない場合が多く見られた。

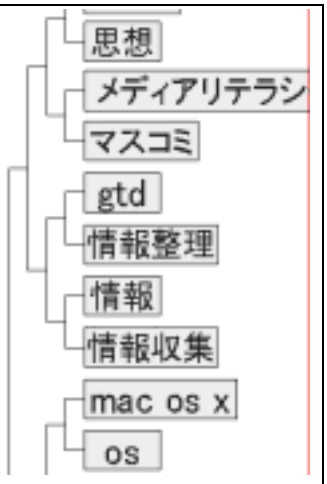

図 3：クラスタの一部

クラスタリングをする際にクラスタ間の距離を算出する手法は複数あり、それぞれの算出方法によって結果として与えられるクラスタ構造は変わってくる。本手法では、最短距離法・最長距離法・重心法・中央値法・ウォード法・可変法[2]の結果をそれぞれ得る。得られた 6 種類のクラスタ構造から、タグからタグへ至る分岐の数が 5 以内である周囲 4〜6 個の特に近しいタグのみを取り出し、類似のタグであるとして併合した。

併合の際、ある手法では類似タグとして取り出されるが別の手法では取り出されない場合、タグ間の分岐を仮に最大より 2 大きい 7 とし、分岐数を加算しタグ同士の非類似度を求めた。

4. 評価実験

大学生 8 人が提示されたランダムなタグと 5〜25 個程度の類似のタグとの関連性を 3 回ずつ評価した。

タグの関連性は「関連がある、関連があると感じる、わからない、関連がないと感じる、関連がない」という 5 項目で回答を得た。関連があると感じる、または関連がないと感じるではタグとタグとの関連性があると感じるが、具体的には示せない場合に選択した。

ランダムなタグと類似タグ一つとの関連性の回答数を合計したものが表 1 だが、タグとして用いられた言葉の意味が理解できない場合も「わからない」が選択されたため、わからないとされた回答については結果に含めていない。

表 1： タグと類似タグ同士の関連性

関連がある	関連があると感じる	関連がないと感じる	関連がない
51	57	44	7

5. 考察

評価実験の結果では「関連がある」と「関連がない」の差が大きく、ある程度、関連性の高いタグ同士を凝縮して整理することができたと思われる。一方、「関連があると感じる」と「関連がないと感じる」の項目についてはあまり明確な差が出ていないが、「興味深い」、「使える」、「これはひどい」のような主観的なタグでは、被験者は用いられた元の文脈を知らないため、タグ単体を見た際に持つイメージとタグの付けられた情報群を俯瞰したイメージとが食い違うためだと考えられる。

主観に基づいて挿入されているタグをフィルタリングすることができれば、さらに正確なタグの関連構造や類似性が見えてくると考えられる。

参考文献

[1] “はてなブックマーク,” http://b.hatena.ne.jp/

[2] 神嶌 敏弘, "クラスタリングとは(クラスター分析とは)," Toshihiro KAMISHIMA, (オンライン) http://www.kamishima.net/jp/clustering/, 2006-12-07.